A Novel Temporal Attentive-Pooling based Convolutional Recurrent Architecture for Acoustic Signal Enhancement

Tassadaq Hussain, Wei-Chien Wang, Mandar Gogate, Kia Dashtipour, Yu Tsao, *Senior Member, IEEE*, Xugang Lu, Adeel Ahsan, and Amir Hussain

***Abstract*—Removing background noise from acoustic observations to obtain clean signals is an important research topic regarding numerous real acoustic applications. Owing to their strong model capacity in function mapping, deep neural network-based algorithms have been successfully applied in target signal enhancement in acoustic applications. As most target signals carry semantic information encoded in a hierarchal structure in short- and long-term contexts, noise may distort such structures nonuniformly. In most deep neural network-based algorithms, such local and global effects are not explicitly considered in a modeling architecture for signal enhancement. In this article, we propose a temporal attentive pooling (TAP) mechanism combined with a conventional convolutional recurrent neural network (CRNN) model, called TAP-CRNN, which explicitly considers both global and local information for acoustic signal enhancement (ASE). In the TAP-CRNN model, we first use a convolution layer to extract local information from acoustic signals and a recurrent neural network (RNN) architecture to characterize temporal contextual information. Second, we exploit a novel attention mechanism to contextually process salient regions of noisy signals. We evaluate the proposed ASE system using an infant cry dataset. The experimental results confirm the effectiveness of the proposed TAP-CRNN, compared with related deep neural network models, and demonstrate that the proposed TAP-CRNN can more effectively reduce noise components from infant cry signals with unseen background noises at different signal-to-noise levels. We further tested the TAP-CRNN ASE system on a downstream infant cry detection (ICD) system, which determines whether a sound segment is involved in an infant cry event. Experimental results show that TAP-CRNN ASE can effectively reduce the noise components, thereby improving the performance of ICD under noisy conditions.**

***Impact Statement*—Recently proposed deep learning solutions have proven useful in overcoming certain limitations of conventional acoustic signal enhancement (ASE) tasks. However, the performance of these approaches under real acoustic conditions is not always satisfactory. In this article, we investigated the use of attention models for ASE. To the best of our knowledge, this is the first attempt to successfully employ a convolutional recurrent neural network (CRNN) with a temporal attentive pooling (TAP) algorithm for the ASE task. The proposed TAP-CRNN framework can practically benefit the assistive communication technology industry, such as the manufacture of hearing aid devices for the elderly and students. In addition, the derived algorithm can benefit other signal processing applications, such as soundscape information retrieval, sound environment analysis in smart homes, and automatic speech/speaker/language recognition systems.**

***Index Terms*—Acoustic signal enhancement (ASE), convolutional neural networks, recurrent neural networks (RNN), bidirectional long-short term memory.**

I. Introduction

Acoustic signals are often distorted owing to additive and convolutional noise, or recording device constraints, which limit the performance of real-world applications, such as soundscape information retrieval [1]–[3], sound environment analysis in a smart home [4]–[6], physiological sound recognition [7]–[10], speaker recognition and verification [11]–[14], automatic speech recognition (ASR) [15]–[18], hearing aids [19], [20], and cochlear implants [21], [22]. The goal of acoustic signal enhancement (ASE) is to suppress the interfering noise signals by minimizing unwanted distortions and transforming noisy input signals into desired clean signals. Several ASE approaches have been proposed in the literature to alleviate background noise problems. However, enhancing the performance of ASE in real-world acoustic environments remains a challenging task. Traditionally, we assume that the target signals and noise follow specific distributions, and thus, a gain function can be estimated to attenuate the noise components. Notable examples include the minimum-mean square error based algorithm [23]–[25] and the Wiener filter [26]. For such approaches, a noise

Manuscript received 15 September 2021; revised 29 December 2021, 4 January 2022, and 13 March 2022; accepted 16 April 2022. Date of publication 25 April 2022; date of current version 17 October 2022. The work of Tassadaq Hussain, Mandar Gogate, Adeel Ahsan, and Amir Hussain were supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/T021063/1. This article was recommended for publication by Associate Editor E. Cambria upon evaluation of the reviewers' comments. *(Corresponding author: Amir Hussain.)*

Tassadaq Hussain, Mandar Gogate, Kia Dashtipour, and Amir Hussain are with the School of Computing, Edinburgh Napier University, EH11 4BN Edinburgh, U.K. (e-mail: tassadaq.hussain@gmail.com; m.gogate@napier.ac.uk; k.dashtipour@napier.ac.uk; a.hussain@napier.ac.uk).

Wei-Chien Wang is with the Institute of Computer and Communication Engineering, National Cheng Kung University, Tainan 701, Taiwan (e-mail: weichian0920@gmail.com).

Yu Tsao is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan. He is also a jointly appointed professor of the Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan 32023, Taiwan (e-mail: yu.tsao@citi.sinica.edu.tw).

Xugang Lu is with the National Institute of Information and Communications Technology, Tokyo 187-0021, Japan (e-mail: xugang.lu@nict.go.jp).

Adeel Ahsan is with the School of Mathematics and Computer Science, University of Wolverhampton, WV1 1LY Wolverhampton, U.K. (e-mail: ahsan.adeel@deepci.org).

Digital Object Identifier 10.1109/TAI.2022.3169995

estimation method is usually required to compute the statistics of the noise signals. Well-known noise estimation approaches include minimum statistics [27], minima controlled recursive averaging (MCRA) [28], and improved MCRA [29].

Another group of traditional ASE algorithms is the subspace-based techniques that split a noisy acoustic signal into two subspaces, one for the clean acoustic signal and another for the noise components, and subsequently, suppress the noise to reconstruct a clean acoustic signal. A notable subspace technique is based on singular value decomposition [30]. The third group of traditional ASE approaches is known as model-based techniques that derive a mathematical model based on human speech production and predict model parameters to perform noise reduction. Successful examples include the harmonic model [31], [32], linear prediction model [33]–[35], and hidden Markov model [36], [37]. Later, matrix decomposition-based methods such as nonnegative matrix factorization [38], [39] and sparse coding were proposed [40], [41] for ASE. These approaches prepare dictionaries of the target signals and noise during the training stage. The noisy signals are decomposed in the online stage; then, the clean signals can be reconstructed based on the prepared matrices with the corresponding activation metrics.

Recently, deep learning (DL)-based approaches have emerged and greatly succeeded in ASE applications. In DL-based architectures, a nonlinear mapping function is estimated to transform noisy acoustic signals to clean ones. For example, in [42], a weighted neural-network-based architecture was proposed for an online ASE system. The proposed system utilized a recurrent neural network (RNN) to enhance short-time Fourier transform (STFT) spectra in a frame-by-frame manner. Furthermore, the authors proposed two additional objectives to enable separately controlling of the importance of acoustic signal distortion and noise reduction. A deep neural network (DNN) architecture was also employed by Zhong *et al.* [43] for single- and multichannel ASE and ASR to predict the real and imaginary components of a signal using noisy and reverberant signals. The proposed framework exhibited good performance in terms of speech quality, intelligibility, and word error rate. Similarly, in [44], a compressed DNN was proposed to conduct real-time ASE and meet the low-latency requirements without compromising the quality and intelligibility of the enhanced signal. Apart from conventional single-stage DL-based models, two-stage temporal convolutional modules have been utilized by Li *et al.* [45] to deal with low signal-to-noise ratios (SNRs). More specifically, the magnitude is estimated in the first stage, which is then combined with the noisy phase to provide a complex spectrum estimation. A secondary network is used as the post-processing module in the second stage to improve the previous estimation, where the residual noise is further suppressed, and the phase information is successfully modified. In the second step, the global residual connection approach is used to increase the training convergence rate. Although deep neural architectures have shown great success for ASE applications, there are still rooms to further improve their achievable performance. A notable one is to further utilize the temporal information of the target and noise signals. In most ASE systems, average pooling or max-pooling is used for temporal feature aggregation. However, determining what features from which temporal regions contribute to ASE serves as an important factor and should be more effectively utilized.

Attention mechanisms have recently been adopted for acoustic signal processing tasks and have shown excellent performance through selectively focusing on the segments of the target signal. The attention model can identify salient regions in acoustic signals to achieve effective performance for classification and regression tasks. In [46], Li *et al.* adopted a dynamic attention method with recursive learning for ASE applications. In [47], the authors proposed a multitask learning strategy along with a multihead self-attention framework to analyze the dependencies between speech and noise signals. In [48], Roy *et al.* investigated a multihead attention network to estimate linear prediction coefficients for clean and noisy speech signals. Convolutional neural network (CNN)- and generative adversarial network (GAN)-based architectures with attention mechanisms for end-to-end ASE applications are employed in [49]–[52].

In [53], [54], Zhao *et al.* and Tan and Wang utilized convolutional RNN (CRNN)-based data-driven models to exploit local structures in both frequency and temporal domains and demonstrated exceptional generalization performance compared to existing DL methods. Inspired by CRNN, where the features are exploited in both frequency and temporal domains, we propose a CRNN with a temporal attentive pooling (TAP) algorithm (aptly termed TAP-CRNN) for the ASE task. In the proposed TAP-CRNN, the convolutional layers extract representative acoustic features and RNN characterizes the long-term temporal information. Meanwhile, the temporal attention mechanism allocates significant segments to effectively train the enhancement model. The developed TAP-CRNN system was evaluated with an infant cry enhancement task. For comparison purposes, we tested the performance of four existing artificial neural network (NN) models, DNN, CNN, RNN (long short-term memory (LSTM)), and CRNN. Experimental results first show that the proposed TAP-CRNN can yield better performance than DNN models in terms of standard ASE evaluation metrics. Moreover, it is verified that TAP-CRNN can effectively enhance a downstream task [infant cry detection (ICD)] in a noisy condition.

The main contributions of this study are as follows:

1) a novel attention algorithm (TAP) is proposed to provide an attention mechanism on top of the existing CRNN framework for the ASE task;
2) the proposed TAP-CRNN provided good discriminative power for target signal restoration and noise reduction compared to CRNN;
3) TAP-CRNN demonstrated superior performance compared to CRNN in terms of four standardized metrics, namely, signal to interference ratio (SIR), signal to artifacts ratio (SAR), segmental SNR (SSNR), and signal distortion ratio (SDR), for the ASE task.

The rest of this article is organized as follows. Section II provides a literature review. Section III discusses the CRNN architecture and the TAP algorithm. The experimental setup and findings are presented in Section IV. Finally, Section V concludes this article.

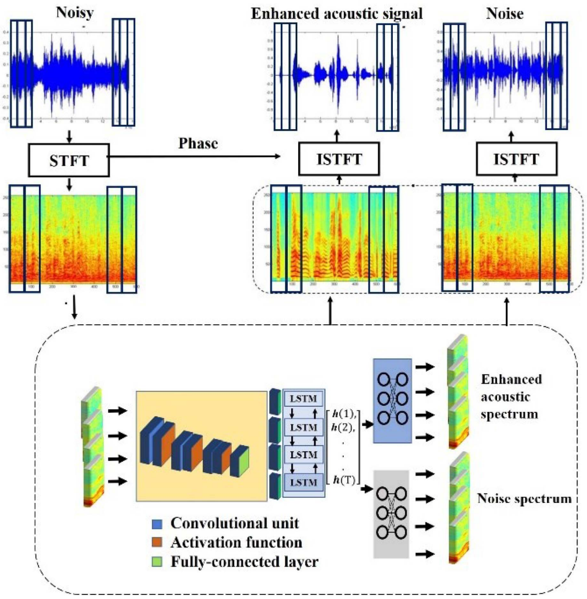


Fig. 1. CRNN-based ASE framework.

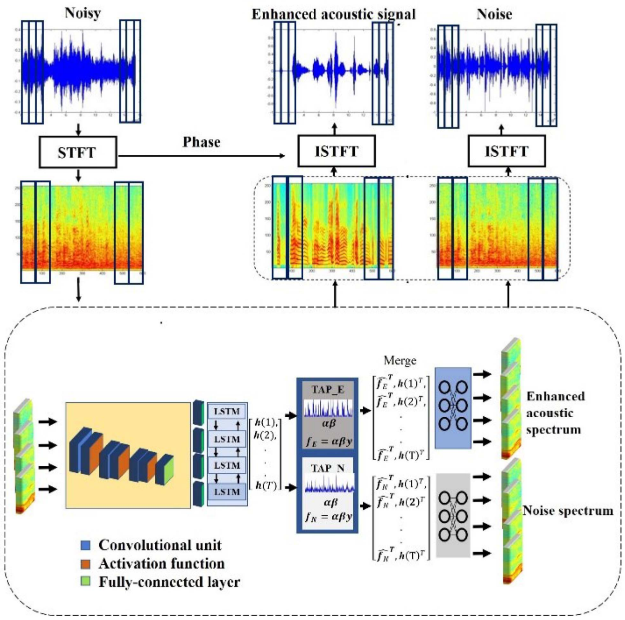


Fig. 2. Proposed TAP-CRNN-based ASE system.

II. Related

This section first introduces conventional DL-based NN models for the ASE task. Next, we briefly discuss the CRNN model followed by the proposed TAP algorithm.

A. Deep Learning Models for ASE

In a DL-based ASE system, a deep neural network model, such as DNN, CNN, or RNN, is used as a mapping function that aims to transform noisy acoustic signals into enhanced ones. In many ASE systems, the noisy and clean acoustic signals are first transformed into spectral features via a STFT for frame-based processing. Then additional processes, such as taking amplitude, logarithm, and band-pass filtering, are applied to the spectral features. The overall architecture of the CRNN-based ASE system is shown in Fig. 1. The goal of the ASE system is to minimize the reconstruction error between the estimated and clean acoustic signals. In the testing stage, the spectral features are fed into the trained ASE model to obtain enhanced spectral features. Finally, the enhanced spectral along with the phase of the original noisy acoustic are converted to the time domain through inverse STFT (ISTFT).

Recent ASE approaches focus on three directions to improve performance. The first direction is to use more suitable objective functions to train the NN models. Traditionally, L2 and L1 distances between the enhanced and clean reference signals are used as objective functions. However, several prior works have shown that the L2 and L1 distances may not be optimal choices. Because ASE is generally used as a pre-processing task, when training the model, we should consider enabling the downstream task to achieve better performance. For example, in assistive speaking applications, it is important to consider quality and intelligibility. Therefore, speech quality- or intelligibility-oriented objective functions should be adopted. Accordingly, Fu *et al.* [55], [56] proposed to directly optimize an NN model with a speech-intelligibility-based objective function, namely short-time objective intelligibility. In [57], a reinforcement learning approach was adopted to optimize the model parameters based on the speech quality index, namely the perceptual evaluation of speech quality (PESQ). Another study derived an objective function that approximated the PESQ function to train the enhancement model [58]. More recently, MetricGAN [59], [60] was proposed to adopt a GAN to enforce the enhancement model to achieve desirable metric scores. The second direction is to use a better model architecture. Examples of successful approaches include models with complex parameters [61]–[63], ensemble learning [64]–[66], dual path [67]–[69], and dual branch [70] architectures. The third direction is to incorporate complementary information from other modalities into ASE applications. Effective modalities include visual data [71]–[74], bone-conductive speech [75], and text information [76], [77].

III. Proposed ASE System

Fig. 2 presents an overview of the proposed TAP-CRNN-based ASE framework. The proposed framework has an additional TAP mechanism compared to the CRNN-based ASE framework. The pseudocode is presented in Algorithm I. The proposed TAP-CRNN system first converts time-domain noisy signals to log power spectral (LPS) features. Then, the LPS features are processed by the ASE module to suppress noise components and accordingly enhance the acoustic signals. Compared to the CRNN-based system, the TAP algorithm computes the attentive contexts from the output of the CRNN, which is then fed to the individual fully connected layers to obtain enhanced acoustic features. Finally, the enhanced LPS features along with the phase information are converted back to time-domain signals using ISTFT.

In the proposed model, the first part is the conventional CNN that takes the LPS as an input to extract spatial-temporal information. Given a sequence of input vectors, $\boldsymbol{X} = \{\boldsymbol{x}(1), .., \boldsymbol{x}(T)\}$ (with *T* frames), the respective outputs of the CNN and RNN blocks can be expressed as

$$\begin{aligned}\boldsymbol{Y} &= \{\boldsymbol{y}\,(1)\,, .., \boldsymbol{y}\,(T)\}\\ \boldsymbol{H} &= \{\boldsymbol{h}\,(1)\,, .., \boldsymbol{h}\,(T)\}\ ,\end{aligned} \tag{1}$$

Subsequently, a bidirectional LSTM (BLSTM) is used to explore forward and backward temporal contextual information and extract global feature representations. During this stage, the output of the ASE module is fed to the TAP algorithm to obtain attentive context information. The attentive context information is then combined with the CRNN's output for calculating enhanced acoustic and noise signals. The combined input, that is, the CRNN output with attentive context, is further processed by the fully connected layers (DNN) to obtain the enhanced acoustic and noise LPS features.

During the online stage, we first calculate the LPS features and phase components of the input noisy acoustic signals. The noisy LPS features are subsequently processed by the CRNN model parameters. Then, the CRNN output is processed by the TAP algorithm to obtain the attentive context of the enhanced spectral features. The enhanced spectral features are forwarded to the fully connected layers for further processing. Finally, the corresponding enhanced LPS features are converted back to time-domain enhanced signal and noise signals using ISTFT, as demonstrated in Fig. 2.

A. TAP Algorithm

In this section, we describe the TAP mechanism in detail. The goal of the TAP mechanism is to introduce attention blocks to the CRNN output by focusing on the salient regions. To maximize the performance of the TAP mechanism, we employed two alternative attention approaches: local and global. The model focuses equally on all regions as part of the global attention. However, when generating an enhanced acoustic signal, the model focuses only on limited local regions. The purpose of global attention is to consider all CNN outputs, as well as the RNN's temporal summarization outputs. A previous study confirmed the benefits of local and global attention integrations on a heart sound classification (regression) task [8]. In this study, the local and global attention integrations on an ASE (regression) task are investigated for the first time. The CRNN with a TAP mechanism is presented in Fig. 3. We use a simple concatenated layer to create the global attentive vector ***c(t)*** for TAP by integrating information obtained from the CNN's output $\boldsymbol{y}(t)$ and RNN's output as follows:

$$\boldsymbol{c}\,(t) = \begin{bmatrix} \boldsymbol{W}_c \boldsymbol{y}\,(t) \\ \boldsymbol{W}_r \boldsymbol{h}\,(T) \end{bmatrix} \tag{2}$$

where $\boldsymbol{h}(t)$ is the temporal sequence, $\boldsymbol{W}_c \in R^{cc \times cnn_{dim}}$, and $\boldsymbol{W}_r \in R^{cr \times rnn_{dim}}$ are the matrices used to reduce $\boldsymbol{y}(t)$ and $\boldsymbol{h}(T)$ dimension; cnn_{dim} and rnn_{dim} are the dimensions of the convolutional and LSTM layers, respectively. cc and cr are the dimensions for convolutional and LSTM parts in $\boldsymbol{c}(t)$, respectively.

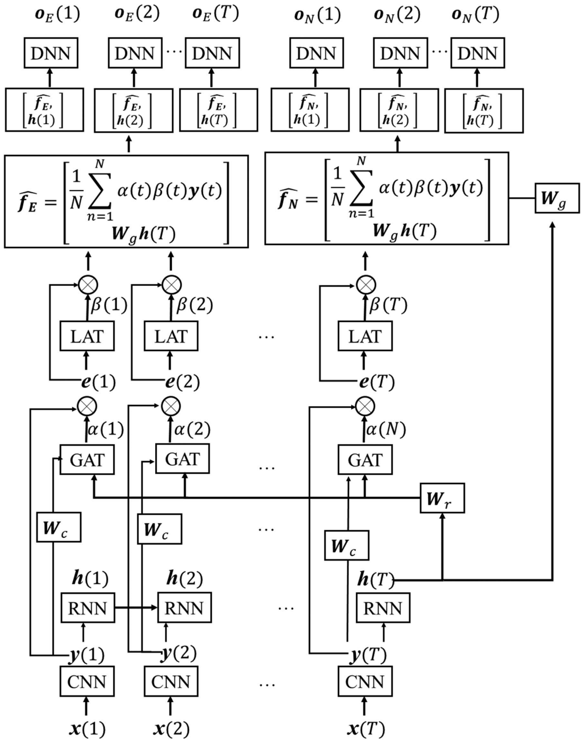

$o_E(1)$
$o_E(2)$
$o_E(T)$
$o_N(1)$
$o_N(2)$
$o_N(T)$
DNN
DNN
DNN
DNN
DNN
DNN
$\widehat{f_{E'}}$ $h(1)$
$\widehat{f_{E'}}$ $h(2)$
$\widehat{f_{E'}}$ $h(T)$
$\widehat{f_{N'}}$ $h(1)$
$\widehat{f_{N'}}$ $h(2)$
$\widehat{f_{N'}}$ $h(T)$
$\widehat{f_E} = \begin{bmatrix}\frac{1}{N}\sum_{n=1}^{N}\alpha(t)\beta(t)y(t)\\ W_g h(T)\end{bmatrix}$
$\widehat{f_N} = \begin{bmatrix}\frac{1}{N}\sum_{n=1}^{N}\alpha(t)\beta(t)y(t)\\ W_g h(T)\end{bmatrix}$
W_g
$\beta(1)$
$\beta(2)$
$\beta(T)$
LAT
LAT
LAT
$e(1)$
$e(2)$
$e(T)$
$\alpha(1)$
$\alpha(2)$
$\alpha(N)$
GAT
GAT
GAT
W_r
W_c
W_c
W_c
$h(1)$
$h(2)$
$h(T)$
RNN
RNN
RNN
$y(1)$
$y(2)$
$y(T)$
CNN
CNN
CNN
$x(1)$
$x(2)$
$x(T)$

Fig. 3. TAP-CRNN model architecture.

The global attentive vector $\boldsymbol{c}(t)$ is subsequently fed to the softmax layer to produce global attention weights $\alpha(t)$ (scalar), which is expressed as

$$\alpha\ (t) = \ \text{softmax}\left(\boldsymbol{u}^T \text{tan}h\ \left(\boldsymbol{c}\,(t) + \boldsymbol{b}_{\text{global}}\right)\right). \tag{3}$$

Here, $\boldsymbol{u} \in R^{(cc+cr)\times 1}$ is the vector used to calculate the global attention weight matrix shared by all-time steps and $\boldsymbol{b}_{\text{global}} \in R^{(cc+cr)\times 1}$ is the global bias matrix. The global attention weights are used to weight the local features obtained from the CNN at each time step as follows

$$\boldsymbol{e}\,(t) = \ \alpha\,(t)\ \boldsymbol{y}\,(t)\,. \tag{4}$$

Apart from global attention, we employed local attention to refine the feature extraction, which is calculated as

$$\beta\,(t) \ = \ \text{softmax}\left(\boldsymbol{v}^T \text{tan}h\left(\boldsymbol{W}_{\text{local}} \boldsymbol{e}\,(t) + \ \boldsymbol{b}_{\text{local}}\right)\right) \tag{5}$$

where $\boldsymbol{W}_{\text{local}} \in R^{l \times cnn_{\text{dim}}}$, $\boldsymbol{b}_{\text{local}} \in R^{l\times 1}$, and $\boldsymbol{v} \in R^{l\times 1}$ are the matrices used to calculate the local attention weight. These weights for local attention are used as feature weights as follows:

$$\boldsymbol{f}\,(t) = \ \alpha\,(t)\,\beta\,(t)\ \boldsymbol{y}\,(t)\,. \tag{6}$$

Here, $\beta(t)$ denotes the local attention output weight vector. The final attentive context is calculated as an average of the weighted outputs and concatenated with the RNN's output $\boldsymbol{h}(T)$ as follows:

$$\hat{\boldsymbol{f}} = \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\alpha\,(t)\,\beta\,(t)\,\boldsymbol{y}\,(t) \\ \boldsymbol{W}_g h\,(T) \end{bmatrix} \tag{7}$$

where $\boldsymbol{W}_g \in R^{rnn_{\text{dim}} \times rnn_{\text{dim}}}$ are the parameter matrices used to concatenate these two vectors. Next, we concatenate the attentive context $\hat{\boldsymbol{f}}(t)$ of the enhanced clean signal with the output of the CRNN $\boldsymbol{h}(t)$ as the input $\boldsymbol{r}(t)$ of the output layers as follows:

$$\boldsymbol{r}(t) = \begin{bmatrix} \hat{\boldsymbol{f}} \\ \boldsymbol{h}(t) \end{bmatrix}. \tag{8}$$

Subsequently, we obtain two $\boldsymbol{r}(t)$ values using TAP for separated noise and enhanced acoustic sound, respectively. The output layers are fully connected layers, where the relationship between the input $\boldsymbol{r}(t)$ and output of the first hidden layer can be expressed as

$$a_1(t) = \sigma(\boldsymbol{W}_1 \boldsymbol{r}(t) + \boldsymbol{b}_1). \tag{9}$$

Here, $\sigma(.)$ is the activation function, $\boldsymbol{W}_1$ is the input weight vector, and $\boldsymbol{b}_1$ is the bias vector of the first hidden layer. Similarly, the relationship for the qth hidden layer can be expressed as

$$\boldsymbol{a}_q(t) = \sigma(\boldsymbol{W}_q \boldsymbol{a}_{q-1}(t) + \boldsymbol{b}_q), \; q = 2, \ldots, Q \tag{10}$$

where Q represents the total number of neurons in the output layer. As a result, the link between the regression and output layers can be formulated as follows:

$$\boldsymbol{o}(t) = G(\boldsymbol{a}_Q(t)) \tag{11}$$

where $G(.)$ is a linear function, and $\boldsymbol{o}(t) \in R^{lps_{\text{dim}} \times 1}$, $t = 1, \ldots, T$.

IV. Experiments

This section first describes the experimental setup and evaluation, then, describes the experimental results. In this study, we used infant cries as the target signals. Various noise types and SNR conditions were sampled to prepare the training and testing sets. As stated in [78], [79], infant cries possess both short-and long-term context structures. The proposed TAP-CRNN can be suitably applied to enhance infant cry signals in noisy conditions.

A. Experimental Setup and Evaluations

The experiments were conducted using an infant cry dataset collected from five infants. For the training set, 400 infant cry utterances were randomly selected and corrupted with six noise types (babble, pink, female speech, male speech, background music, and cocktail party) at three SNR levels (−5, 0, and 5 dB) to generate 400 (utterances) × 6 (noise types) × 3 (SNRs) = 7200 training utterances. Moreover, 100 infant cry utterances (different from those used in the training set) were selected to form the test set. We intentionally created a training-test mismatch scenario, where the test utterances were formed by contaminating 100 clean infant cry utterances with two background noise signals (i.e., a fan and a piece of background music) at four mismatch SNR levels (i.e., −6, −2, 2, and 6 dB) to generate 100 (utterances) × 2 (noise types) × 4 (SNRs) = 800 test infant cry utterances. In this study, we used the 257-dimensional LPS as the acoustic feature.

To evaluate the performance of the proposed ASE system, we adopted two standardized evaluation metrics: SSNR and SDR. The SSNR measures the average SNR values over short acoustic signal segments (15–20 ms). A higher SSNR value indicates that the ASE method can more effectively suppress background noise. The SDR measures the distortion of the enhanced signal, where a lower SDR indicates less distortion in the enhanced signal. In contrast to standardized evaluation metrics, which focus on the enhanced acoustic signal, the background noise signals also contain important information. Thus, we also considered the ASE as a monaural source-separation task that aims to separate the target and background noise signals given a noisy signal. To evaluate such a task, we adopted two additional evaluation metrics, namely, SIR and SAR, to estimate the performance. All of the ASE systems reported in this study were implemented on TensorFlow, using an RMSprop optimizer with a learning rate of 1e–5 and a batch size of 32. The experiments were run on a single NVIDIA GeForce GTX 1080 Ti GPU with 11 GB of memory.

Algorithm: TAP-CRNN.

```
Input: Noisy acoustic signal sequence $\boldsymbol{X}$ $\{\boldsymbol{x}(1),..,\boldsymbol{x}(T)\}$,
       where $\boldsymbol{x}(t) \in R^{257\times 1}$.
Output: Separated target signal $\boldsymbol{O}_E = \{\boldsymbol{o}_E(1),..,\boldsymbol{o}_E(T)\}$
and Separated noise signal $\boldsymbol{O}_N = \{\boldsymbol{o}_N(1),..,\boldsymbol{o}_N(T)\}$,
where $\boldsymbol{o}_E(t) \in R^{257\times 1}$ and $\boldsymbol{o}_N(t) \in R^{257\times 1}$.
for $t = 1$ to$T$ do
    $\boldsymbol{y}(t) = $ convolution$(\boldsymbol{x}(t))$
    $\boldsymbol{h}(t) = $ lstm$(\boldsymbol{y}(t))$
end for

$\boldsymbol{Y} = \{\boldsymbol{y}(1),..,\boldsymbol{y}(T)\}$
$\boldsymbol{H} = \{\boldsymbol{h}(1),..,\boldsymbol{h}(T)\}$
    Calculate global attentive weight $\alpha(t)$
$\alpha(t)$ = GAT_Attention ([$\boldsymbol{y}(t), \boldsymbol{h}(T)$])
for $t = 1$ to$T$ do
    $\boldsymbol{e}(t) = \alpha(t) \cdot \boldsymbol{y}(t)$
end for

$\boldsymbol{E} = \{\boldsymbol{e}(1),..,\boldsymbol{e}(T)\}$
    Calculate local attentive weight $\beta(t)$
$\beta(t)$ = LAT_Attention ([$\boldsymbol{e}(t)$])

for $t = 1$ to$T$ do
$\boldsymbol{f}(t) = \beta(t) \cdot \boldsymbol{e}(t)$
end for

Attentive feature: $\hat{\boldsymbol{f}} = [\begin{matrix} \sum_{t=1}^{T} \boldsymbol{f}(t) \\ \boldsymbol{W}_g \boldsymbol{h}(T) \end{matrix}]$,

for $t = 1$ to$T$ do
    $\boldsymbol{r}(t) = [\begin{matrix} \hat{\boldsymbol{f}} \\ \boldsymbol{h}(t) \end{matrix}]$
    $\boldsymbol{o}_E(t) = $ FullyConnected_layer$(\boldsymbol{r}(t))$
    $\boldsymbol{o}_N(t) = $ FullyConnected_layer$(\boldsymbol{r}(t))$
end for
return $\boldsymbol{O}_E$ and $\boldsymbol{O}_N$
```

TABLE I
PERFORMANCE COMPARISONS OF TAP-CRNN AND CRNN OVER TWO NOISE TYPES (FAN AND BACKGROUND MUSIC) AT FOUR SNR LEVELS

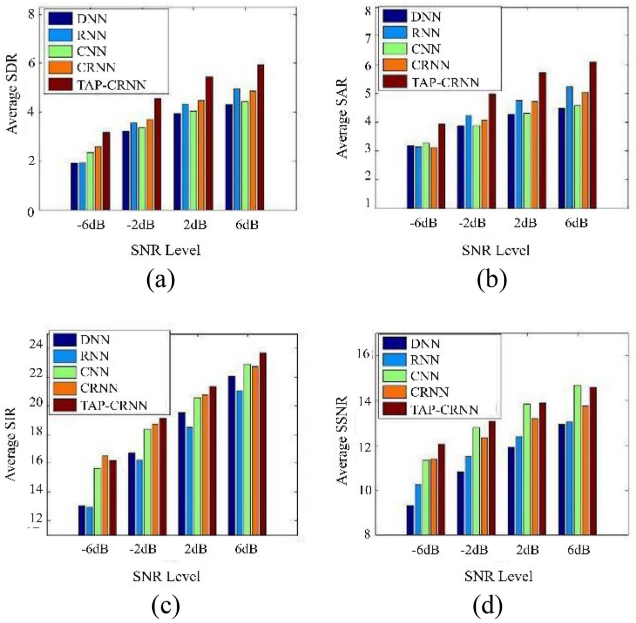
	CRNN				**TAP-CRNN**			
	SDR	**SIR**	**SAR**	**SSNR**	**SDR**	**SIR**	**SAR**	**SSNR**
-6dB	3.29	18.3	3.69	11.83	4.40	19.29	4.74	12.54
-2dB	4.30	20.9	4.55	12.43	5.10	21.56	5.34	13.00
2dB	4.78	22.66	4.89	13.41	5.51	23.45	5.66	14.04
6dB	4.86	24.45	5.05	13.59	5.70	25.24	5.79	14.32
Avg.	4.31	21.58	4.55	12.82	5.18	22.39	5.38	13.48

B. Evaluation of TAP Mechanism

In this section, we assessed the performance of the TAP mechanism by comparing the results of the TAP-CRNN and CRNN frameworks. We tested the performance with stationary (fan noise) and nonstationary noise (background music) types, both of which were not involved in the training data. Table I reports the enhancement results in terms of average SDR, SIR, SAR, and SSNR scores for the three noise types at four SNR levels (−6, −2, 2, and 6 dB). The average scores across the four SNR levels are listed in the table as well. From the table, we can note that TAP-CRNN outperforms CRNN consistently over different SNR levels in all evaluation metrics, where the average improvement of SDR is 20.20% (from 4.31 to 5.18), that of SIR is 3.74% (from 21.58% to 22.39%), that of SAR is 18.43% (from 4.55% to 5.38), and that of SSNR is 5.15% (from 12.82 to 13.48%). These results confirmed the effectiveness of the TAP mechanism.

C. Comparison to Different Neural Architectures

Next, we compared the performance of the TAP-CRNN framework to that of four well-known DL-based ASE frameworks, namely, DNN, CNN, RNN, and CRNN. In the DNN framework, we used six fully connected layers, each having 512 units. The CNN model had two convolutional layers, each having 32 filters with both kernel size of 3 and stride of 2, and three fully-connected layers, with each layer containing 512 units. The RNN framework is formed by two BLSTM layers, each with 256 units, followed by two fully connected RNN layers, each having 256 neurons. In the CRNN framework, the input noisy LPS features were first processed by the CNN at each time step and subsequently fed to the two layers of BLSTM, each having 128 units. The output of the BLSTM is later concatenated and processed by the two independent DNNs, each having two hidden layers with 128 units. TAP-CRNN has additional TAP blocks, as described in Section III-A. Each block of the TAP consists of an individual DNN, which has two feedforward layers with a size of 256 units. One of the two subsequent blocks of TAP is for ASE, where the enhanced acoustic signal is attained by the DNN. The other block of TAP is used to recover the noise signal. The overall architecture of the TAP-CRNN is shown in Fig. 2. The output of the subsequent layers of CRNN is processed by two independent TAP blocks (DNNs) to reconstruct the enhanced acoustic signal by separating the noise signal.



(a) (b) (c) (d)

Fig. 4. Evaluation results for different DNNs on the infant cry dataset. (a) SDR. (b) SAR. (c) SIR. (d) SSNR.

Fig. 4 presents a summary of the average SDR, SIR, SAR, and SSNR scores yielded by the TAP-CRNN against four related ASE systems at different SNR levels under stationary and non-stationary noise conditions. For all performance evaluation metrics, namely noise reduction (SSNR) and target-signal-noise separation (SDR, SIR, and SAR), a high score indicates better ASE performance. As shown in Fig. 4, DNN and RNN exhibited similar behaviors for SDR, SRI, and SAR metrics at low SNR levels. Meanwhile, RNN achieved higher SDR, SAR, and SSNR scores compared to the DNN for higher SNR levels. The results suggest although DNN that uses a fully-connected neural architecture can well enhance acoustic signals, RNN, which incorporates temporal information, can more effectively separate noise components from the noisy input. The CNN models, on the other hand, have been proved to be good at characterizing the spatial context when performing acoustic enhancement [54], [55]. The results in Fig. 4 confirmed that CNN outperforms RNN and DNN in terms of SIR and SSNR while underperforming RNN in terms of SDR and SAR. The inconsistent results may come from different properties of CNN and RNN models. Finally, CRNN is designed to combine the advantages of CNN and RNN,

TABLE II
COMPUTATION COST (IN TERMS OF FLOPS) AND MODEL SIZE (IN TERMS OF # PARAMETERS) OF CRNN AND TAP-CRNN

	FLOPs (M)	#Parameters (M)
CRNN	3.58	9.51
TAP-CRNN	5.39	11.03

where CNN captures spatial information and RNN models the temporal characteristics of speech features. For most evaluation metrics, CRNN yields better performance than individual CNN and RNN models across different SNR levels.

The TAP mechanism is adopted to further disentangle blended information learned by the CRNN; in this way, CRNN may focus on the significant frame by adopting global and local attention weights and deciding where to focus attention when generating the enhanced acoustic signal and noise. In Fig. 4, the results of TAP-CRNN demonstrated better performance as compared to the other ASE systems in almost all performance evaluation metrics. The results confirm the effectiveness of the TAP mechanism that further enhances the ASE capability on top of CRNN. Table II shows the online computation cost in terms of floating-point operations (FLOPs) and model complexity in terms of the number of model parameters of the CRNN and TAP-CRNN, where FLOPs represent the computation cost of multiply and accumulate (MAC) operations. From the table, we note that TAP-CRNN only moderately increases the computational computation cost and model complexity compared to CRNN.

D. Spectrogram Analyses

In addition to quantitative analyses based on the evaluation metrics, we intend to visually compare the ASE capabilities of TAP-CRNN to other DL-based models. Fig. 5 displays the spectrogram plots of the enhanced acoustic signals achieved by DNN, RNN, CNN, CRNN, and TAP-CRNN. Fig. 5(a) and (b) demonstrates the spectrogram plots of noisy acoustic signal (termed as Noisy) corrupted with non-stationary noise (i.e., *male speech*) at -6 dB SNR and clean infant cry acoustic signal (termed as Clean) for comparison. From Fig. 5, we observe that all ASE models successfully suppressed the background noise components while dealing with non-stationary noise and effectively restoring the low-frequency acoustic regions at very low SNR levels (-6 dB), as indicated by the rectangular boxes. In Fig. 5(f), we can note that the CRNN effectively removed the background noise from the noisy signal; however, it misjudged the region in the rectangular box as an acoustic region, which shows that it has learned blended information. However, TAP-CRNN restored the low-frequency acoustic regions shown in the rectangular box closer to the clean utterance spectrogram plot by helping the CRNN focus on the significant frame using the global and local attention weights when generating enhanced acoustic signals.

To investigate the full effect of the TAP mechanism, Fig. 6 demonstrates how the weights of the sequence generated by the TAP mechanism. The displayed visualization belongs to an infant cry utterance, which has been contaminated with nonstationary noise. As shown in Fig. 6, the TAP learns alignments that correspond very strongly to the target acoustic signals by focusing on and paying attention to the significant frames. The attention mechanism correctly identifies and restores the frames by discriminating noise and acoustic signal components through assigning high weights to the salient regions or frames in noisy signals.

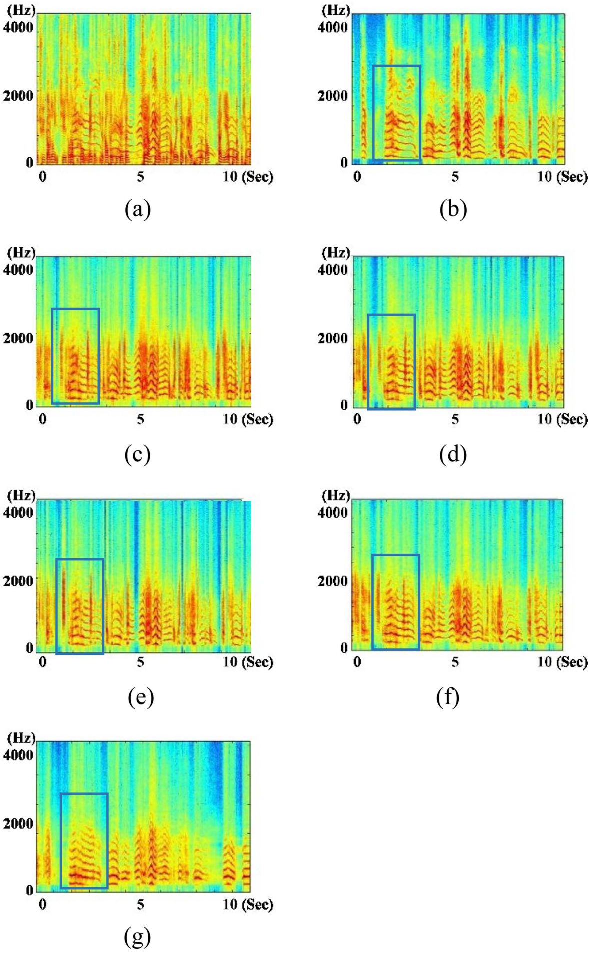


Fig. 5. Spectrogram plots of enhanced acoustic signals processed by (a) Noisy, (b) Clean, (c) DNN, (d) RNN, (e) CNN, (f) CRNN, and (g) TAP-CRNN. The unprocessed signal (a) and the corresponding clean signal (b) are also presented for comparison.

E. ASE for Robust ICD

We further test the ASE system using a downstream task: ICD. The ICD task aims to determine whether a sound segment involves an infant cry event or not. In this set of experiments, we used a training set from ESC-50 [84] to establish an ICD system. The training data contained 228 sound segments of two classes, 108 segments involved infant cry events, and 120 segments involved background environment sounds (no infant cry). The testing set included 500 sound segments, in which 250 segments involved infant cry events, and 250 segments involved background sounds only (no infant cry event). These testing segments were contaminated with five types of noises (pink

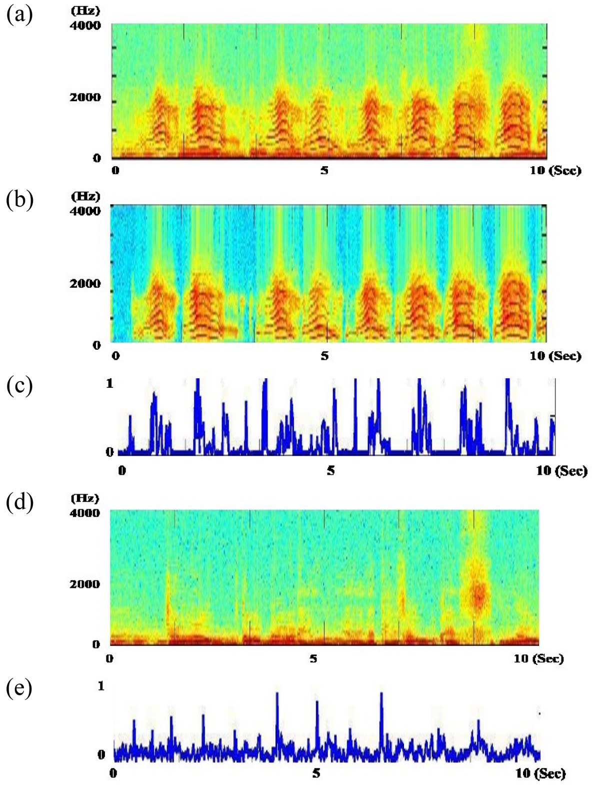


Fig. 6. (a) Spectrogram plot of a noisy signal. (b) Spectrogram plot of the enhanced acoustic signal. (c) Attention weights of TAP_E. (d) Spectrogram plot of the noise signal. (e) Attention weights of TAP_N (computed by $\alpha(t)\beta(t)$).

TABLE III
ASSESSMENT MATRIX

	Ground Truth	
Predicted Class	TP (true positive)	FP (false positive)
	FN (false negative)	TN (true negative)

noise, bubble noise, two female mixed speech, and one male speech) at varying SNR levels (from −2 to 2 dB). Each segment was of a 1-s length. The ICD system was built based on temporal convolutional residual networks (TC-ResNet8), which had been used for the Google small foot-print KeyWord-Spotting task [85]. Since the ICD is a detection task, typical metrics for pattern detection tasks were used for evaluation, including accuracy, precision, recall, and F1-score. Table III shows the four items used in the evaluations.

Equations (12)–(15) show the definitions of four metrics that we used to evaluate the ICD performance

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{12}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{13}$$

$$\mathrm{F1} \ = \ 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{14}$$

TABLE IV
ACCURACY, F1-SCORE, PRECISION, AND RECALL RESULTS OF THE ICD SYSTEMS WITHOUT ASE AND WITH TAP-CRNN ASE

	Accuracy	F1-score	Precision	Recall
Without ASE	78.4	0.749	0.894	0.644
With ASE	89.0	0.879	0.976	0.800

and

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}. \tag{15}$$

Among the four metrics, recall is the true positive rate, and precision represents the positive predictive rate. The F1-score is the harmonic mean of recall and precision. Accuracy reflects the classification results.

Table IV shows the ICD system without ASE (without ASE) and with TAP-CRNN-based ASE (with ASE) in terms of accuracy, F1-score, precision, and recall. For without ASE, the testing segments were directly sent to the pretrained ICD system to test recognition; for ASE, the testing segments were first processed by TAP-CRNN for ASE and then sent to the pretrained ICD system to test recognition. From Table IV, TAP-CRNN can effectively improve the classification results under noisy conditions, confirming TAP-CRNN's outstanding capability to improve the downstream ICD task.

V. CONCLUSION

In this article, we investigated using attention models for the ASE task. To the best of our knowledge, this is the first attempt to employ a TAP mechanism on a CRNN for such applications. The results highlighted the unequal importance of the segments of each frame for ASE. The proposed TAP mechanism considers local and global regions of a sequence, extracts significant features from regions of the sequence, and extracts significant features from each frame. The global attention mechanism identifies salient parts of the sequence, whereas the local attention mechanism minimizes the attention error. Comparative experiments indicate that the TAP-CRNN possesses positive discriminative capabilities for target signal restoration and noise reduction compared to the CRNN model. The effectiveness of the proposed TAP-CRNN mechanism is demonstrated under mismatched nonstationary environments at severe SNRs compared to other deep-learning ASE frameworks as well. Experimental results also confirmed the effectiveness of TAP-CRNN-based ASE for improving a downstream ICD task under noisy conditions. Note that the performance of the proposed TAP-CRNN was analyzed (trained and tested) using a relatively limited infant cry dataset. In future research, we aim to consider more diverse training data and different nonstationary noise types to further assess the performance of TAP-CRNN. Moreover, we will explore noise-aware and SNR-aware-based training for TAP-CRNN to further enhance system performance in real-time environments. Finally, we will build on our recent research [80]–[83] and extend the TAP-CRNN model by investigating a compressed version of TAP-CRNN to contextually exploit multimodal information, such as audio-visual lip-reading, to meet strict latency constraints and generalized performance requirements for next-generation multimodal hearing aids.

Acknowledgment

The authors are very grateful to the anonymous reviewers for their insightful comments and suggestions which have helped improve the quality of this paper.

References

[1] R. Bardeli, D. Wolff, F. Kurth, M. Koch, K.-H. Tauchert, and K.-H. Frommolt, “Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring,” *Pattern Recogn. Lett.*, vol. 31, pp. 1524–1534, 2010.

[2] B. Pijanowski *et al.*, “Soundscape ecology: The science of sound in the landscape,” *Bio. Sci.*, vol. 61, no. 3, pp. 203–216, 2011.

[3] T. H. Lin and Y. Tsao, “Source separation in ecoacoustics: A roadmap towards versatile soundscape information retrieval,” *Remote Sens. Ecol. Conservation*, vol. 6, no. 3, pp. 236–247, 2020.

[4] S. Krstulović, “Audio event recognition in the smart home,” in *Computational Analysis of Sound Scenes Events*. Berlin, Germany: Springer, pp. 335–371, 2018.

[5] M. A. Sehili *et al.*, “Sound environment analysis in smart home,” in *Proc. Int. Joint Conf. Ambient Intell.*, 2012, pp. 208–223.

[6] B. D. Coensel and D. Botteldooren, “Smart sound monitoring for sound event detection and characterization,” in *Proc. Int. Congr. Noise Control Eng.*, 2014, pp. 3442–3451.

[7] K. H. Tsai *et al.*, “Blind monaural source separation on heart and lung sounds based on periodic-coded deep autoencoder,” *IEEE J. Biomed. Health Informat.*, vol. 24, no. 11, pp. 3203–3214, Nov. 2020.

[8] J.-K. Wang *et al.*, “Automatic recognition of murmurs of ventricular septal defect using convolutional recurrent neural networks with temporal attentive pooling,” *Sci. Rep.*, vol. 10, no. 21797, pp. 1–10, 2020.

[9] S. H. Fang *et al.*, “Detection of pathological voice using cepstrum vectors: A deep learning approach,” *J. Voice*, vol. 33, no. 5, pp. 634–641, 2019.

[10] L. Shi, K. Du, C. Zhang, H. Ma, and W. Yan, “Lung sound recognition algorithm based on VGGish-BiGRU,” *IEEE Access*, vol. 7, pp. 139438–139449, Sep. 2019.

[11] S. Kataria, P. S. Nidadavolu, J. Villalba, and N. Dehak, “Analysis of deep feature loss based enhancement for speaker verification,” in *Proc. Odyssey*, 2020, pp. 459–466.

[12] Y. Jung, Y. Choi, H. Lim, and H. Kim, “A unified deep learning framework for short-duration speaker verification in adverse environments,” *IEEE Access*, vol. 8, pp. 175448–175466, Sep. 2020.

[13] D. Cai, W. Cai, and M. Li, “Within-sample variability-invariant loss for robust speaker recognition under noisy environments,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6469–6473.

[14] S. Shon and J. Glass, “Multimodal association for speaker verification,” in *Proc. Interspeech*, 2020, pp. 2247–2251.

[15] R. Li, X. Wang, S. H. Mallidi, S. Watanabe, T. Hori, and H. Hermansky, “Multi-stream end-to-end speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process*, vol. 28, pp. 646–655, Dec. 2020.

[16] S. Wang, W. Li, S. M. Siniscalchi, and C. Lee, “A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6219–6223.

[17] C. Donahue, B. Li, and R. Prabhavalkar, “Exploring speech enhancement with generative adversarial networks for robust speech recognition,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5024–5028.

[18] Y.-H. Tu, J. Du, and C.-H. Lee, “Speech enhancement based on teacher-student deep learning using improved speech presence probability for noise-robust speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2080–2209, Dec. 2019.

[19] N. Shankar, G. S. Bhat, and I. M. S. Panahi, “Efficient two-microphone speech enhancement using basic recurrent neural network cell for hearing and hearing aids,” *J. Acoustical Soc. Amer.*, vol. 148, no. 1, pp. 389–400, 2020.

[20] I. Fedorov *et al.*, “TinyLSTMs: Efficient neural speech enhancement for hearing aids,” in *Proc. Interspeech*, 2020, pp. 4054–4058.

[21] R.-Y. Tseng, T.-W. Wang, S.-W. Fu, C.-Y. Lee, and Y. Tsao, “A study of joint effect on denoising techniques and visual cues to improve speech intelligibility in cochlear implant simulation,” *IEEE Trans. Cogn. Devlop. Syst.*, vol. 13, no. 4, pp. 984–994, Dec. 2021.

[22] N. Mamun, S. Khorram, and J. H. L. Hansen, “Convolutional neural network-based speech enhancement for cochlear implant recipients,” in *Proc. Interspeech*, 2019, pp. 4265–4269.

[23] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Jun. 1984.

[24] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Feb. 1985.

[25] Y. Lu and P. C. Loizou, “A geometric approach to spectral subtraction,” *Speech Commun.*, vol. 50, no. 6, pp. 453–466, 2008.

[26] P. Scalart, “Speech enhancement based on a priori signal to noise estimation,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 1996, vol. 2, pp. 629–632.

[27] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, May 2001.

[28] I. Cohen and B. Berdugo, “Noise estimation by minima controlled recursive averaging for robust speech enhancement,” *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[29] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[30] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen, “Reduction of broad-band noise in speech by truncated QSVD,” *IEEE Trans. Speech Audio Proces.*, vol. 3, no. 6, pp. 439–448, Nov. 1995.

[31] J. S. Lim and A. V. Oppenheim, “Enhancement and bandwidth compression of noisy speech,” *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[32] J. S. Lim, *Speech Enhancement.* Englewood Cliffs, NJ, USA: Prentice-Hall, 1983.

[33] K. K. Paliwal and A. Basu, “A speech enhancement method based on Kalman filtering,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 1987, pp. 177–180.

[34] J. D. Gibson, B. Koo, and S. D. Gray, “Filtering of colored noise for speech enhancement and coding,” *IEEE Trans. Signal Process.*, vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[35] S. Gannot, D. Burshtein, and E. Weinstein, “Iterative and sequential kalman filter-based speech enhancement algorithms,” *IEEE Trans. Speech Audio Process.*, vol. 6, no. 6, pp. 373–385, Jul. 1998.

[36] Y. Ephraim, D. Malah, and B.-H. Juang, “On the application of hidden Markov models for enhancing noisy speech,” *IEEE Trans. Acoust. Speech Signal Process.,* vol. 37, no. 12, pp. 1846–1856, Dec. 1989.

[37] Y. Ephraim, “A Bayesian estimation approach for speech enhancement using hidden Markov models,” *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[38] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, “Speech denoising using nonnegative matrix factorization with priors,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 4029–4032.

[39] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, 2013.

[40] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, “Wind noise reduction using non-negative sparse coding,” in *Proc. Workshop Mach. Learn. Signal Process.*, 2007, pp. 431–436.

[41] C. D. Sigg, T. Dikk, and J. M. Buhmann, “Speech enhancement with sparse coding in learned dictionaries,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4758–4761.

[42] X. Yangyang, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, “Weighted speech distortion losses for neural-network-based real-time speech enhancement,” in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 871–875.

[43] Z.-Q. Wang, P. Wang, and D. Wang, “Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR,” *IEEE/ACM Trans. Audio, Speech, Lang. Proces.*, vol. 28, pp. 1778–1787, May 2020.

[44] K. Tan and D. Wang, “Towards model compression for deep learning based speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1785–1794, May 2021.

[45] A. W. L. Li, C. Zheng, C. Fan, and X. Li, “Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1829–1843, May 2021.

[46] A. Li, C. Zheng, C. Fan, R. Peng, and X. Li, “A recursive network with dynamic attention for monaural speech enhancement,” in *Proc. Interspeech*, 2020, pp. 2422–2426.

[47] Y. Koizumi, K. Yatabe, M. Delcroix, Y. Masuyama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 181–185.

[48] S. K. Roy, A. Nicolson, and K. K. Paliwal, "DeepLPC-MHANet: Multi-head self-attention for augmented Kalman filter-based speech enhancement," *IEEE Access*, vol. 9, pp. 70516–70530, May 2021.

[49] X. Xiang, X. Zhang, and H. Chen, "A convolutional network with multi-scale and attention mechanisms for End-to-End single-channel speech enhancement," *IEEE Signal Process. Lett.*, vol. 28, pp. 1455–1459, Jun. 2021.

[50] A. Pandey and D. Wang, "Dense CNN with self-attention for time-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1270–1279, Mar. 2021.

[51] H. Phan *et al.*, "Self-attention generative adversarial network for speech enhancement," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7103–7107.

[52] C. H. Yang, J. Qi, P. Y. Chen, X. Ma, and C. H. Lee, "Characterizing speech adversarial examples using self-attention u-net enhancement," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3107–3111.

[53] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional recurrent neural networks for speech enhancement," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2401–2405.

[54] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Proc. Interspeech*, 2018, pp. 3229–3233.

[55] S. W. Fu, T. W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[56] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 825–838, Jan. 2020.

[57] Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, and Y. Haneda, "DNN-based source enhancement to increase objective sound quality assessment score," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1780–1792, Oct. 2018.

[58] J. M. Martín-Doñas, A. M. Gomez, J. A. Gonzalez, and A. M. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1680–1684, Nov. 2018.

[59] S.-W. Fu, C. F. Liao, Y. Tsao, and S. D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. 36th Int. Conf. Mach. Learn*, 2019, pp. 2031–2041.

[60] S.-W. Fu *et al.*, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, 2021, pp. 201–205.

[61] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," 2019, *arXiv:1903.03107*.

[62] Y. Hu, Y. Liu, S. Lv, M. Xing, Y. F. S.Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2021, pp. 2472–2476.

[63] Z. Q. Wang, P. Wang, and D. Wang, "Complex spectral mapping for single- and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1778–1787, 2020.

[64] C. Yu *et al.*, "Speech enhancement based on denoising autoencoder with multi-branched encoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2756–2769, Oct. 2020.

[65] S. E. Chazan, J. Goldberger, and S. Gannot, "Deep recurrent mixture of experts for speech enhancement," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, 2017, pp. 359–363.

[66] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 5, pp. 967–977, May 2016.

[67] A. Pandey and D. Wang, "Dual-path self-attention RNN for real-time speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 46–50.

[68] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement," in *Proc. Interspeech*, 2021, pp. 2811–2815.

[69] K. Zhang, S. He, H. Li, and X. Zhang, "DBNet: A dual-branch network architecture processing on spectrum and waveform for single-channel speech enhancement," in *Proc. Interspeech*, 2021, pp. 2821–2825.

[70] G. Yu, A. Li, Y. Wang, Y. Guo, H. Wang, and C. Zheng, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7847–7851.

[71] J. C. Hou, S. S. Wang, Y. H. Lai, Y. Tsao, H. W. Chang, and H. M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, 2018.

[72] S. Y. Chuang, H. M. Wang, and Y. Tsao, "Improved lite audio-visual speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, no. 30, pp. 1345–1359, Feb. 2022.

[73] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[74] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, I. Girin, and R. Horaud, "The conversation: Deep audio-visual speech enhancement Audio-visual speech enhancement using conditional variational auto-encoders. The conversation: Deep audio-visual speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1788–1800, 2020.

[75] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, "Time-domain multi-modal bone/air conducted speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 1035–1039, Jun. 2020.

[76] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Proc., 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1760–1764.

[77] C.-F. Liao, Y. Tsao, X. Lu, and H. Kawai, "Incorporating symbolic sequential modeling for speech enhancement," in *Proc. Interspeech*, Sep. 2019, pp. 2733–2737.

[78] S. Ntalampiras, "Audio pattern recognition of baby crying sound events," *J. Audio Eng. Soc.*, vol. 63, no. 5, pp. 358–369, 2015.

[79] P. S. Zeskind, S. Parker-Price, and R. G. Barr, "Rhythmic organization of the sound of infant crying," *Devlop. Psychobiol., J. Int. Soc. Devlop. Psychobiol.*, vol. 26, no. 6, pp. 321–333, 1993.

[80] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, "CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement," *Inf. Fusion*, vol. 63, pp. 273–285, 2020.

[81] M. Gogate, K. Dashtipour, P. Bell, and A. Hussain, "Deep neural network driven binaural audio-visual speech separation," in *Proc. Int. Joint Conf. Neural Networks*, 2020, pp. 1–7.

[82] A. Adeel, M. Gogate, and A. Hussain, "Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments," *Inf. Fusion*, vol. 59, pp. 163–170, 2020.

[83] M. Gogate, K. Dashtipour, and A. Hussain, "Visual speech in real noisy environments (VISION): A novel benchmark dataset and deep learning-based baseline system," in *Proc. Interspeech*, 2020, pp. 4521–4525.

[84] K. J. Piczak, "ESC: Dataset for environmental sound classification," *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.

[85] S. Choi *et al.*, "Temporal convolution for real-time keyword spotting on mobile devices," in *Proc. Interspeech*, 2019, pp. 3372–3376.